A Search and Discovery Tool - AMASE

Cynthia Y. Cheung

NASA Goddard Space Flight Center, Astrophysics Data Facility,
Greenbelt, MD 20771

Nick Roussopoulos, Stephen Kelley

University of Maryland, Institute of Advanced Computer Studies

James Blackwell

Raytheon STX Corporation

Abstract.
The Astrophysics Multispectral Archive Search Engine *(AMASE)* is a metadatabase prototype built using object-oriented database methodology which allows mission data to be searched easily by scientific parameters. We discuss the implementation issues that arise in integrating heterogeneous data sources: the form and contents of metadata; the creation of cross-system object identifiers; specific tools to handle bulk loading and incremental data loading; interoperability with relational databases; and our future work to extend the capabilities of *AMASE.*

1. Introduction

The importance of multi-spectral research in solving fundamental problems in astrophysics is undeniable. It is generally advantageous, for example, to use multi-wavelength observations to interpret a class of celestial objects. However, NASA archival data are usually not organized to support multi-spectral research. The data are products of space missions and are stored in the archives according to mission-specific parameters, in disparate mass storage systems and in different formats. This makes it quite challenging to locate data for research that cut across mission or spectral boundaries. The Astrophysics Multispectral Archive Search Engine *(AMASE)* aims at providing a uniform multi-mission and multi-spectral interface, enabling *global* searches of these heterogeneous, distributed data holdings using astronomical classification and complex spatial queries. Users are able to locate relevant mission data for research without detailed knowledge of the missions beforehand. The URL for the *AMASE* homepage is `http://amase.gsfc.nasa.gov/`.

2. The *AMASE* Design & Implementation

AMASE uses object-oriented data base (OODB) techniques to "merge" astronomical data from heterogeneous sources. These techniques *encapsulate* the existing data, metadata, and associated documentation and bibliography into an abstract *Astronomical_Object*. The flexibility and modeling capability of OODB allow for diversity and conplexity within a common framework. The relationship between objects can be captured in the database schema and implemented at the time of data loading. The *Astronomical_Object* can be characterized by very complex and rich data types, and can take on multiple classification to reflect both its intrinsic scientific nature and the instrument parameters connected with scientific data acquisition. The *Astronomical_Object* acts as a 'directory' pointing users to, for example, the catalogs in the Astronomical Data Center containing information about the celestial object and the mission data in the distributed archives with different observatory-instrument settings. When a user queries for science data, the search criteria are issued against the scientific parameters of the *Astronomical_Object* while the response is given in terms of mission data descriptors. The mission data descriptors identify the relevant mission data sets and give the archive location and retrieval information. In this way, *AMASE* provides a scientific view into the distributed data archives without changing the existing underlying mission-oriented structure.

AMASE is implemented using the commercial object-relational product *Informix-Illustra DBMS*. It offers the dual advantages of an OODB and a relational database, with complex and user-defined data types that are not normally supported in a relational database, a standard SQL non-procedural query language, the robustness of a storage manager, a query optimizer, and the maturity of the technology provided by relational systems.

3. Building the Knowledge Base

To support queries using scientific parameters, *AMASE* must first build an astronomical "knowledge base". The domain knowledge needed to describe modern astronomical classification structure is captured from experts in the different disciplines. The heritage and relationships of different mission data products are likewise captured from the appropriate mission scientists. Once the basic classification schema is defined, the knowledge base is populated with valid astronomical measurements. These are generally available in published astronomical catalogs and can be captured by automated procedures. But, naturally, the data in the knowledge base will be subject to revision in the course of scientific inquiry: new data will be acquired from future missions, new object cross–identifications will be made, resulting in a better understanding of physical phenomena, and possibly, a new object classification scheme. So a mechanism for incremental buildup of knowledge and subsequent knowledge evolution is built into the *AMASE* architecture, enabling it to become a true search and discovery engine.

Loading an OODB is much more involved than loading a relational database. Correlations between objects are created and maintained by the database *at load time*. Linkages are represented by object identifiers (OIDs) in the database. We choose to use common astronomical names as the cross-catalog and cross-mission OIDs in *AMASE*. The dual use of astronomical names to refer to celestial objects as well as database objects provides a friendly user query mechanism. The OIDs are often not available when the load file is written, because the detected sources of an observation are not yet identified. In such cases, the semi-automated bulk loading procedure will create *incomplete* objects with surrogate identifier when observations are first loaded, then *complete* the object by adding correct pointers to the relevant astronomical characteristics when the target is identified. We are building incremental bulk loading utilities and user utilities to support updates of object characteristics, to "merge" objects upon positive cross-identification, and to separate objects into components when they are resolved by high-resolution observations.

4. Accessing Remote Databases

The major archive sites of the NASA astrophysics missions are geographically distributed. Their data holdings are frequently updated as new observations are taken and new data are released to the public archive. *AMASE* must provide access to data at these archive sites and maintain current information about their data holdings. The schema of each remote archive database is first mapped to the *AMASE* schema. Then a small but sufficient amount of metadata is ingested to create a "partially materialized" view (Roussopoulos et al. 1993) in the *AMASE* database. An alternative is to create "virtual" object in *AMASE* that is "pointer-based" and populated with references to the data values and locations in the remote archives (Roussopoulos 1982). Batch updates of these derived views are carried out by a special utility built using the Sybase Gateway of the *Informix-Illustra DBMS.*

5. Query Capabilities

There are two main ways to query the *AMASE* database: by object name and by position. The query by astronomical name illustrates the advantage of an OODB implementation. Since all information that pertains to a requested object is linked at load time, the execution is very efficient. The positional information is returned as part of the search result and is not used at all during the search.

To efficiently process queries based on spatial position, we build spatial indexes on every database object that contains spatial attributes. We use $packedR-trees$ (Roussopoulos & Leifker 1985) to significantly reduce the search selection set and are extending their use to support the spatial ordering of objects and data discovery based on object position.

Since there is no "natural" ordering of multi-attribute data, especially spatial data, results are returned in the order they are retrieved from the index. To obtain the result based on "closeness" to the query object, most query implementation will select the data first, write the results to a temporary object, then sort them for presentation. But the cost in time and space for this approach will be prohibitive for queries with a large number of results. To address this issue, we are developing a "K-Nearest Neighbor" access method that will generate and return the K objects closest to a given spatial position in order (Roussopoulos et al. 1995).

6. Future Work – Data Discovery Utilities

A principal goal of *AMASE* is to support astrophysics research by providing a position-based "data mining" facility. The first implementation of this facility will support field-of-view (FOV) correlation. We would use an instrument FOV to scan through the *AMASE* database to extract information for off-target objects, correlate them to other data (from other catalogs/observations) via spatial (primary) and other attributes, then present the results to a scientist for review.

Other data discovery utilities include view-caching and user-defined types. Query results can be cached into the database and used for repetitive queries and/or refinement. An astronomer can build his/her own profile of interested objects from the database. Results cached in views can be accessed much more rapidly as they are compact and require a much narrower search. To support user-defined classes, the system must have a utility to rearrange the class hierarchy and its implied inheritance, and to assign object identifiers to newly derived class object. We plan to use "materialized views" as the mechanism for supporting user-defined types (Roussopoulos 1991, Roussopoulos et al. 1993). Part of this work was funded by an earlier NASA AISR grant.

7. Summary

The *AMASE* prototype is currently populated with selected data products from six NASA astrophysics missions, as well as various astronomical catalogs. Users can search for information by specifying either object name and/or various positional information. The search can be further qualified by selecting the object type, the spectral region and the mission. Sorting and visualization capabilities will soon be available through the ADC Data Viewer ([T]2.2). In the coming year, we shall provide the capabilities to interoperate with other data archives and develop utilities for user-defined views, positional data mining, and object cross identification.

References

[T]2.2 this volume

Roussopoulos, N. 1982, *ACM Transactions on Database Systems*, 7(2), 258

Roussopoulos, N. & Leifker, D. 1985, *Procs. of 1985 ACM SIGMOD Intl. Conf. on Management of Data, Austin, 1985*, 17

Roussopoulos, N. 1991,*ACM–Transactions on Database Systems*, 16(3), 535

Roussopoulos, N. , Economou, N. and Stamenas, A. 1993, *IEEE Trans. on Knowledge and Data Engineering*, 5(5), 762

Roussopoulos, N. Kelley, S. & Vincent, S. 1995, *Proc. of ACM SIGMOD, San Jose, California, May 22-25, 1995*, 71